

# Deep Supervision with Shape Concepts for Occlusion-Aware 3D Object Parsing

Chi Li<sup>1</sup>, M. Zeeshan Zia<sup>2</sup>, Quoc-Huy Tran<sup>2</sup>, Xiang Yu<sup>2</sup>, Gregory D. Hager<sup>1</sup>, and Manmohan Chandraker<sup>2</sup>

<sup>1</sup>Johns Hopkins University

<sup>2</sup>NEC Laboratories America, Inc.

## Abstract

Monocular 3D object parsing is highly desirable in various scenarios including occlusion reasoning and holistic scene interpretation. We present a deep convolutional neural network (CNN) architecture to localize object semantic parts in 2D image and 3D space while inferring their visibility states, given a single RGB image. Our key insight is to exploit domain knowledge to regularize the network by deeply supervising its hidden layers, in order to sequentially infer a causal sequence of intermediate concepts associated with the final task. To acquire training data in desired quantities with ground truth 3D shape and intermediate concepts, we render 3D object CAD models to generate large-scale synthetic data and simulate challenging occlusion configurations between objects. We train the network only on synthetic data and demonstrate state-of-the-art performances on real image benchmarks including an extended version of KITTI, PASCAL VOC, PASCAL3D+ and IKEA for 2D and 3D key-point localization and instance segmentation. The empirical results substantiate the utility of deep supervision scheme by demonstrating effective transfer of knowledge from synthetic data to real images, resulting in less overfitting compared to standard end-to-end training.

## 1. Introduction

The world around us is rich in structural regularity, particularly when we consider man-made objects such as cars or furniture. Studies in perception show that the human visual system imposes structure to reason about stimuli[34]. Consequently, early work in computer vision studied perceptual organization as a fundamental precept for recognition and reconstruction [22, 23]. However, algorithms designed on these principles relied on hand-crafted features (such as corners or edges) and hard-coded rules (such as junctions or parallelism) to hierarchically reason about [26] or learn [32] abstract concepts such as shape. Such approaches suffered from limitations in the face of real-world complexities. In contrast, with the advent of convolutional neural networks

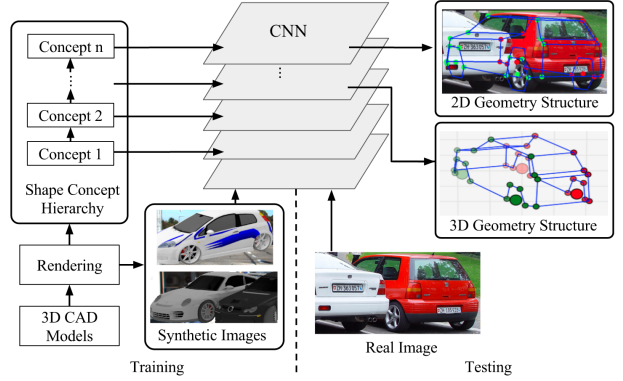


Figure 1: Overview of our approach. We use synthetic training images with intermediate shape concepts to deeply supervise the hidden layers of a CNN. At test time, given a single real image of an object, we demonstrate accurate localization of semantic parts in 2D and 3D, while being robust to intra-class appearance variations as well as occlusions.

(CNNs) in recent years, there has been tremendous progress in end-to-end trainable feature learning for abstract tasks such as object detection, segmentation and reconstruction.

In this paper, we posit that it is advantageous to consider a middle ground, where abstract shape concepts such as pose, visibility or 3D structure can be used to better supervise features learned in an end-to-end trainable CNN. In particular, we combine such early intuitions on shape concepts with the discriminative power of modern CNNs to parse 3D object geometry across intra-class appearance variations, including complex phenomena such as occlusions. Specifically, we demonstrate that imposing appropriate supervision for shape concepts such as pose or visibility that constitute 3D shape, within the intermediate layers of a CNN, allows greater accuracy in estimating the semantic elements of an object observed in a single image.

To illustrate this, we use the 3D skeleton[37] as a shape representation, where semantically meaningful object parts (such as the wheels of a car) are represented by 3D keypoints and their connections define the 3D structure of an object category. This representation is efficient compared to 3D

volumes [4] or meshes [36, 14, 27, 16, 29] in conveying the semantic information necessary for shape reasoning in applications such as autonomous driving.

We introduce a novel CNN architecture which jointly models multiple shape concepts including object pose, keypoint locations and visibility in Section 3. We first generalize the deep supervision [17] in Section 3.1 and show its better generalization capability than the standard end-to-end training. This motivates our network architecture in Section 3.2 in which we deeply supervise convolutional layers at different depths with intermediate shape concepts. Further, instead of using expensive manual annotations, Section 3.3 proposes to render 3D CAD models to create synthetic images with concept labels and simulate the challenging occlusion configurations for robust occlusion reasoning. Figure 1 introduces our framework and Figure 2 illustrates a particular instance of deeply supervised CNN using shape concepts. We denote our network as “DISCO” short for Deep supervision with Intermediate Shape Concepts.

At test time, our CNN trained on only synthetic images generalizes well to real images. In particular, it outperforms comparable architectures without supervision for intermediate shape concepts. This fact demonstrates the intimacy of shape concepts for 3D object parsing, despite that we ignore aspects of photorealism such as material and illumination in our rendering pipeline. Our experiments also show that deep supervision coupled with shape concepts at various depths is beneficial, compared to imposing supervision for all the concepts at the top layer. In Section 4, we quantitatively demonstrate significant improvements over prior state-of-the-art for 2D keypoint and 3D structure prediction on PASCAL VOC, PASCAL3D+[42], IKEA[20] and an extended KITTI [7] dataset (KITTI-3D).

We note that most existing approaches [14, 16, 40, 45, 47] estimate 3D geometry by comparing projections of parameterized shape models with separately predicted 2D patterns, such as keypoint locations or heat maps. This makes prior methods sensitive to partial view ambiguity [18] and incorrect 2D structure predictions. Moreover, scarce 3D annotations for real images further limit their performances. In contrast, we make several novel contributions as follows to alleviate those concerns:

- We demonstrate the utility of rendered data with access to intermediate shape concepts. In addition, we model occlusions by appropriately rendering multiple object configurations, which presents a novel way of exploiting 3D CAD data for realistic scene interpretation.
- We apply intermediate shape concepts to deeply supervise the layers of a CNN for better generalization than the standard end-to-end training. This enables accurate localization of semantic object parts in 2D and 3D even with occlusion, truncation and large appearance variations.
- Our method achieves the state-of-the-art performance for

2D and 3D semantic part localization on several public benchmarks.

## 2. Related Work

**3D Skeleton Estimation** This class of work models 3D shape as a linear combination of shape bases and optimizes basis coefficients to fit computed 2D patterns such as heat maps [45] or object part locations [47]. The single image 3D interpreter network (3D-INN) [39] presents a sophisticated CNN architecture to estimate a 3D skeleton based only on detected visible 2D joints. The training of 3D-INN is not jointly optimized for 2D and 3D keypoint localization. Further, the decoupling of 3D structure from rich object appearance leads to partial view ambiguity and thus 3D prediction errors.

**3D Reconstruction** A generative inverse graphics model is formulated by [16] for 3D mesh reconstruction by matching mesh proposals to extracted 2D contours. Recently, given a single image, autoencoders have been exploited for 2D image rendering [5], multi-view mesh reconstruction [36] and 3D shape regression under occlusion [27]. The encoder network learns to invert the rendering process to recognize 3D attributes such as object pose. However, methods such as [36, 27] are quantitatively evaluated only on synthetic data and seem to achieve limited generalization to real images. Other works such as [14] formulate an energy-based optimization framework involving appearance, keypoint and normal consistency for dense 3D mesh reconstruction, but require both 2D keypoint and object segmentation annotations on real images for training. Volumetric frameworks with either discriminative [4] or generative [29] modeling infer a 3D shape distribution in voxel grids given one or multiple images of the same object. However, due to the highly redundant nature of voxel grid representations, they are limited to low resolutions up to 32x32x32 for now. Lastly, 3D voxel exemplars [41] jointly recognize the 3D shape and occlusion pattern by template matching, which is not scalable to more object types and complex shapes.

**3D Model Retrieval and Alignment** This line of work estimates 3D object structure by retrieving the closest object CAD model and performing alignment, using 2D images [46, 1, 19, 24, 42] and RGB-D data [2, 10]. Unfortunately, limited number of CAD models can not represent all instances in one object category. Further, the retrieval step is slow for a large CAD dataset and the alignment is sensitive to error in estimated pose.

**Pose Estimation and 2D Keypoint Detection** “Render for CNN” [35] synthesizes 3D CAD model views as additional training data besides real images for object viewpoint estimation [35]. We extend this rendering pipeline to support object keypoint prediction and model occlusion. Viewpoint prediction is utilized in [38] to significantly boost the performance of 2D landmark localization. Recent work such

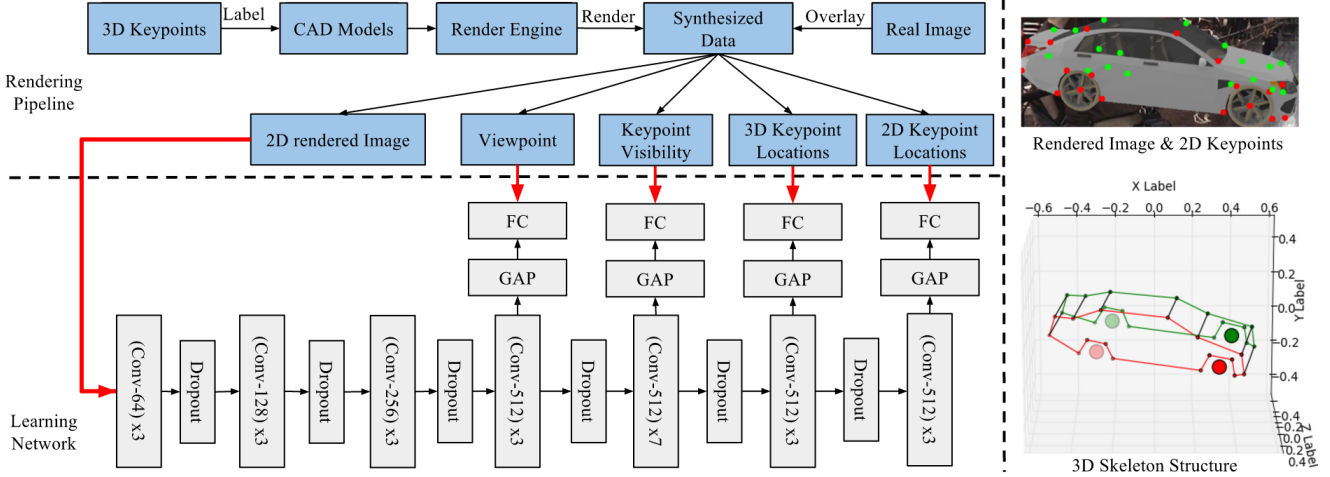


Figure 2: Visualization of our rendering pipeline (top-left), DISCO network (bottom-left), an example of rendered image and its annotations of 2D keypoints (top-right) as well as 3D skeleton (bottom-right).

as DDN [44] optimizes deformation coefficients based on the PCA representation of 2D keypoints to achieve state-of-the-art performance on face and human body. Dense feature matching approaches which exploit top-down object category knowledge [13, 45] also obtain recent successes, but our method yields better results.

### 3. Deep Supervision with Shape Concepts

In the following, we introduce a novel CNN-based framework for 3D shape parsing which incorporates constraints through intermediate shape concepts such as object pose, keypoint locations, and visibility information. More specifically, for each object class, we define a 3D skeleton where joints represent parts, and their connections define object geometry. Our goal is to infer, from a single view (RGB image) of the object, the locations of keypoints in the 2D/3D spaces and their visibility. We discuss notation and motivation of our deep supervision in Section 3.1. Subsequently, we present the network architecture in Section 3.2 which exploits synthetic data generated from the rendering pipeline in Section 3.3.

#### 3.1. Deep Supervision

Our approach draws inspiration from Deeply Supervised Nets (DSN) [17]. However, whereas [17] supervises each layer by the final label to accelerate training convergence, we apply deep supervision, using multiple intermediate concepts intrinsic to the ultimate task, for better generalization.

More concretely, let  $\mathcal{Z} = \{(x, y)\}$  represent the set of all pairs of input  $x$  and labels  $y$  for a particular supervised learning task. The associated optimization problem for a

multi-layer CNN is:

$$W^* = \operatorname{argmin}_W \sum_{(x, y) \in \mathcal{Z}} l(y, f(x, W)) \quad (1)$$

where  $l(\cdot, \cdot)$  is a problem specific loss,  $W = \{W_1, \dots, W_N\}$  stands for the weights of  $N$  layers, and function  $f$  is defined based on the network structure.

In practice, although the empirical optimal solution  $\widehat{W}^*$  minimizes (1) over the population of training data  $\mathcal{Z}$ , this solution may suffer from overfitting. That is, given a new population of data  $\mathcal{Z}'$ , the performance of  $f(\cdot, W)$  on this population might be substantially lower than on  $\mathcal{Z}$ . This is particularly the case when, for example, we train on synthetic data but test on real data.

One way to address the overtraining is through regularization which biases solutions  $W$  toward those which exhibit better generalization. In particular, when using simulated data, it is possible to introduce regularization that biases the network to reproduce physical quantities that are causally related to the final answer – for example, object pose determines object part visibility. Intuitively, the idea is to prefer solutions that reflect the underlying physical structure of the problem which is entangled in the original training set. Since deeper layers in CNNs represent more complex concepts due to growing size of receptive fields and more non-linear transformations stacked along the way, we could realize our intuition by explicitly enforcing hidden layers to reach a sequence of known intermediate concepts with growing complexity towards the final task.

To this end, we define the augmented training set  $\mathcal{A} = \{(x, \{y_1, \dots, y_N\})\}$  with additional supervisory signals  $\{y_1, \dots, y_{N-1}\}$ . We further define  $W_{1:i} = \{W_1, \dots, W_i\}$  as the weights for the first  $i$  layers of the CNN. We also denote

$h_i = f(\cdot, W_{1:i})$  as the internal state of layer  $i$ . We now extend (1) to the additional training signals  $y_i$  by introducing an additional function  $y_i = g(h_i, v_i)$  parameterized by the weight  $v_i$ . Letting  $V = \{v_1, \dots, v_{N-1}\}$ , we can now write a new objective trained over  $\mathcal{A}$ :

$$\widehat{W}^*, \widehat{V}^* = \underset{W, V}{\operatorname{argmin}} \sum_{(x, \{y_i\}) \in \widehat{\mathcal{A}}} \sum_{i=1}^N \lambda_i l_i(y_i, g(f(x, W_{1:i}), v_i)) \quad (2)$$

The above objective can be optimized by simultaneously backpropagating the errors of all supervisory signals scaled by  $\lambda_i$  on each  $l_i$  to  $W_{1:i}$ . From the perspective of the original problem, new constraints through  $y_i$  act as additional regularization on the hidden layers, thus “shaping” the network toward solutions which, as we empirically show, exhibit much better generalization than solutions to (1).

### 3.2. Network Architecture

To set up (2) for 2D/3D semantic parsing, we must first choose a causal ordering of concepts with growing complexity that predict 2D/3D keypoint locations. We have chosen, in order, (1) object viewpoint, which determines (2) keypoint visibility, which induces (3) 3D structure prediction, which finally leads to (4) 2D keypoint prediction for the full set of keypoints including ones that are not visible. We impose this causal sequence to deeply supervise the network at certain depths as shown in Fig. 2 and minimize four intermediate losses  $l_i$  in (2), with other losses removed.

Our network resembles the VGG network [33] and consists of deeply stacked  $3 \times 3$  convolutional layers. Unlike VGG, we remove local spatial pooling and couple each convolutional layer with batch normalization [11] and ReLU, which defines the  $f(x, W_{1:i})$  in (2). This is motivated by the intuition that spatial pooling leads to the loss of spatial information. Further,  $g(h_i, v_i)$  is constructed with one global average pooling (GAP) layer followed by one fully connected (FC) layer with 512 neurons, which is different from stacked FC layers in VGG. In Sec. 4.1, we empirically show that these two changes are critical to significantly improve the performance of VGG like networks for 2D/3D landmark localization.

To further reduce the issue of over-fitting, we deploy dropout [15] layers between the convolutional layers. At layers 4, 8, 12, we perform the downsampling using convolution layers with stride 2. The bottom-left of Fig. 2 illustrates the details of our network architecture. “(Conv-A)xB” means A stacked convolutional layers with filters of size BxB. We deploy 25 convolutional layers in total.

We use L2 loss at all points of supervision and set all loss weights to 1. In practice, we only consider the azimuth angle of the object viewpoint with respect to a canonical pose. We further discretize the azimuth angle into  $M$  bins and regress it to a one-hot encoding (the entry corresponding



Figure 3: Examples of synthesized training images for simulating the multi-car occlusion.

to the predicted discretized pose is set to 1 and all others to 0). Keypoint visibility is also represented by a binary vector with 1 indicating occluded state of a keypoint. 2D keypoint locations are normalized to  $[0, 1]$  with the image size along the width and height dimensions. We center 3D keypoint coordinates of a CAD model at the origin and scale them to set the longest dimension (along X,Y,Z) to unit length. The CAD models are assumed to be aligned along the principal coordinate axes, and registered to the canonical pose, as is the case for ShapeNet [3] dataset. During training, each loss is backpropagated to train the network jointly.

### 3.3. Synthetic Data Generation

Unsurprisingly, our approach needs a large amount of training data because it is based on deep CNNs and involves more fine-grained labels than many other visual tasks such as object detection. Furthermore, we aim for the method to work with occluded test cases. Therefore, we need to generate training examples that are representative of realistic occlusion configurations caused by multiple objects in close proximity and image boundary truncations. To obtain such large scale training data, we extend the data generation pipeline of “Render for CNN” [35] with 2D/3D landmarks and visibility information.

An overview of the rendering process is shown in the upper-left of Fig. 2. We pick a small subset of CAD models from ShapeNet [3] for a given object category and manually annotate 3D keypoints on each CAD model. Next, we render each CAD model via the open-source tool Blender while randomly sampling graphics parameters from a uniform distribution including camera viewpoint, number/strength of light sources, and surface gloss reflection. Finally, we overlay the rendered images on real image backgrounds to avoid over-fitting to synthetic data [35]. We crop the object from each rendered image and extract the object viewpoint, 2D/3D keypoint locations and their visibility states from the render engine as the training labels. In Fig. 2, we show an example of rendering and its 2D/3D annotations.

To model multi-object occlusion, we randomly select two different object instances and place them close to each other without overlapping in 3D space. During rendering, we compute the occlusion ratio of each instance by calculating the fraction of visible 2D area versus the complete 2D projection of CAD model. Keypoint visibility is computed



by ray-tracing. We select instances with occlusion ratios ranging from 0.4 to 0.9. Fig. 3 shows representative training examples where cars are occluded by other nearby cars. For truncation, we randomly select two image boundaries (left, right, top, or bottom) of the object and shift them by  $[0, 0.3]$  of the image size along that dimension.

## 4. Experiment

**Dataset and metrics** We empirically demonstrate competitive or superior performance over several state-of-the-art methods, on a number of public datasets: PASCAL VOC (Sec. 4.2), PASCAL3D+ [42] (Sec. 4.3) and IKEA [20] (Sec. 4.4). In addition, we evaluate our method on KITTI-3D where we generate 3D keypoint annotations on a subset of car images from KITTI dataset [7]. For training, we select 472 cars, 80 sofa and 80 chair CAD models from ShapeNet [3]. Each car model is annotated with 36 keypoints [47] and each sofa or chair model is labeled with 14 keypoints [42]<sup>1</sup>. We synthesize 600k car images including occluded instances and 200k images of fully visible furniture (chair+sofa). We select rendered images of 5 CAD models from each object category as the validation set.

We use PCK and APK metrics [43] to evaluate the accuracy of 2D keypoint localization. A 2D keypoint prediction is correct when it lies within a specified radius  $\alpha * L$  of the ground truth, where  $L$  is the larger dimension of the image with  $0 < \alpha < 1$ . PCK is the percentage of correct keypoint predictions given the object location and keypoint visibility. APK is the mean average precision of keypoint detection computed by associating each estimated keypoint with a confidence score. In our experiments, we use the regressed values of keypoint visibility as confidence scores. We extend 2D PCK and APK metrics to 3D by defining a correct 3D keypoint prediction whose euclidean distance to the ground truth is less than  $\alpha$  in normalized coordinates.

**Training details** We use stochastic gradient descent with momentum 0.9 to train the proposed CNN from scratch. Learning rate starts at 0.01 and decreases by one-tenth when the validation error reaches a plateau. The weight decay is set to 0.0001 and the input image size is 64x64. The network is initialized following [9] and the batch size is 100. For car model training, we form each batch using a mixture of fully visible, truncated and occluded cars, numbering 50, 20 and 30, respectively. For the furniture, each batch consists of 100 images of chair and sofa mixed with random ratios. The network is implemented with Caffe[12]. We use DISCO to name our network.

### 4.1. KITTI-3D

**2D and 3D Structure Prediction** We use 2D keypoint annotations of 2040 KITTI [7] car instances provided by Zia

<sup>1</sup>We use 10 chair keypoints consistent with [39] for evaluation on IKEA.

et al. [47]. To create KITTI-3D, we further label each car image with occlusion type and 3D keypoint locations. We define four occlusion types: no occlusion (or fully visible cars), truncation, multi-car occlusion (the target car is occluded by other cars) and occlusion created by other objects. The number of images for each type is 788, 436, 696 and 120, respectively. To obtain 3D groundtruth, we fit a PCA model trained on the 3D keypoint annotations on CAD data, by minimizing the 2D projection error for the known 2D landmarks. We only provide 3D keypoint labels for fully visible cars because the occluded or truncated cars do not contain enough visible 2D keypoints for precise 3D alignment. We refer the readers to the supplementary material for more details about the 3D annotation and some labeled examples in KITTI-3D.

We compare our method with the recent works DDN [44] and WarpNet [13] for 2D keypoint localization and Zia et al. [47] for 3D structure prediction. We use the source codes of all comparative methods provided by the author. Further, we enhance the WarpNet (denoted as WN-gt-yaw) by using the groundtruth poses of test images to retrieve 30 labeled synthetic car images for landmark transfer. All methods are trained on the same set of synthetic training images and tested on cropped cars using ground truth locations. We can see that DISCO outperforms all baseline methods significantly on all occlusion types. Additionally, we compare DISCO with its several variants. First, we incrementally remove the deep supervisions used in DISCO one by one. DISCO-vis-3D-2D, DISCO-3D-2D and plain-2D are networks without pose, pose+visibility and pose+visibility+3D, respectively. Next, plain-3D is the first 22-layer DISCO network with only 3D supervision. Further, plain-all places all supervision signals on the final convolutional layer. Finally, DISCO-VGG replaces the downsampling and GAP in DISCO with the non-overlapping spatial pooling (2x2) and a fully connected layer with 512 neurons, respectively.

In Table 1, we report PCK accuracies for various methods<sup>2</sup>. We observe that DISCO significantly outperforms competitors in both 2D and 3D keypoint localization. Moreover, by incrementally adding supervisions into DISCO, we observe a monotonically increasing trend of 2D and 3D accuracies among the variants of DISCO: plain-2D or plain-3D < DISCO-3D-2D < DISCO-vis-3D-2D < DISCO. Further, plain-all is superior to plain-2d and plain-3d, while DISCO exceeds plain-all by 4.3% on 2D-All and 2.9% on 3D-Full. These experiments suggest that joint modeling of 3D shape concepts is better than independent modeling. We attribute this success to the complementary nature of our labels and the regularization effect via deep supervision. Finally, DISCO-VGG performs significantly worse than DISCO by 16.0% on 2D-All and 5.6% on 3D-Full,

<sup>2</sup>We cannot report Zia et al.[47] on occlusion categories because only a subset of images has valid results in those classes.

Method	2D					3D	3D-yaw
	Full	Truncation	Multi-Car Occ	Other Occ	All	Full	Full
DDN [44]	67.6	27.2	40.7	45.0	45.1	NA	
WN-gt-yaw* [13]	88.0	76.0	81.0	82.7	82.0	NA	
Zia et al. [47]	73.6	NA				73.5	7.3
plain-2D	88.4	62.6	72.4	71.3	73.7	NA	
plain-3D	NA					90.6	6.5
plain-all	90.8	72.6	78.9	80.2	80.6	92.9	3.9
DISCO-3D-2D	90.1	71.3	79.4	82.0	80.7	94.3	3.1
DISCO-vis-3D-2D	92.3	75.7	81.0	83.4	83.4	95.2	2.3
DISCO-Vgg	83.5	59.4	70.1	63.1	69.0	89.7	6.8
DISCO	93.1	78.5	82.9	85.3	85.0	95.3	2.2

Table 1: PCK[ $\alpha = 0.1$ ] accuracies (%) of different methods for 2D and 3D keypoint localization on KITTI-3D dataset. WN-gt-yaw [13] uses groundtruth pose of the test car. The red color indicates the best result.

PCK[ $\alpha = 0.1$ ]	Long[21]	VKps[38]	DISCO
Full	55.7	81.3	81.8
Full[ $\alpha = 0.2$ ]	NA	88.3	93.4
Occluded	NA	62.8	59.0
Big Image	NA	90.0	87.7
Small Image	NA	67.4	74.3
All [APK $\alpha = 0.1$ ]	NA	40.3	45.4

Table 2: PCK[ $\alpha = 0.1$ ] accuracies (%) of different methods for 2D keypoint localization on the car category of PASCAL VOC. Red color indicates the best result.

which confirms our intuition to remove local spatial pooling and adopt global average pooling.

We also evaluate DISCO on detection bounding boxes computed from RCNN[8] with IoU > 0.7 to the groundtruth of KITTI-3D. The PCK accuracies by DISCO on 2D-All and 3D-Full are 88.3% and 95.5% respectively, which are even better than the ones on groundtruth bounding boxes in Table 1. It can be attributed to the fact that 2D groundtruth locations in KITTI do not tightly bound the object areas because they are only the projections of 3D groundtruth bounding boxes. This result shows that DISCO is robust to the location noises in detection algorithms. We refer readers to more numerical details in the supplementary material.

## 4.2. PASCAL VOC

We evaluate DISCO on the PASCAL VOC 2012 dataset for 2D keypoint localization [43]. Unlike KITTI-3D where car images are captured on real roads and mostly in low resolution, PASCAL VOC contains car images with larger appearance variations and extreme occlusions. In Table 2, we compare our results with state-of-the-art [38, 21] on various sub-classes of the test set: fully visible cars (denoted as “Full”), occluded cars, high-resolution (average size 420x240) and low-resolution images (average size 55x30). Please refer to [38] for details of the test setup.

We observe that DISCO outperforms [38] by 0.6% and 5.1% on PCK at  $\alpha = 0.1$  and  $\alpha = 0.2$ , respectively. In addition, DISCO is robust to low-resolution images, improving 6.9% accuracy on low-resolution set compared with [38]. However, DISCO is inferior on the occluded car class and high-resolution images, attributable to our use of small images (64x64) for training and the fact that our occlusion simulation cannot capture more complex occlusion in typical road scenes. Finally, we compute APK accuracy at  $\alpha = 0.1$  for DISCO on the same detection candidates used in [38]<sup>3</sup>. We can see that DISCO outperforms [38] by 5.1% on the entire car dataset (Full+Occluded). This suggests DISCO is more robust to the noisy detection results and more accurate on keypoint visibility inference than [38]. We attribute this to global structure modeling of DISCO during training where the full set of 2D keypoints teaches the network to resolve the partial view ambiguity.

Note that some definitions of our car keypoints [47] are slightly different from [43]. For example, we annotate the bottom corners of the front windshield but [43] label the side mirrors. In our experiments, we ignore this annotation inconsistency and directly apply the prediction results. Further, unlike [21, 38], we do not use the PASCAL VOC train set, since our intent is to study the impact of deep supervision with shape concepts available through a rendering pipeline. Thus, the better performance is expected when real images with consistent labels are used for training.

## 4.3. PASCAL3D+

PASCAL3D+ [42] provides object viewpoint annotations for PASCAL VOC objects by aligning manually chosen 3D object CAD models onto the visible 2D keypoints. Because only a few CAD models are used for each category, 3D keypoint locations are not accurate. Thus, we use the evaluation metric proposed by [42] which measures the 2D

<sup>3</sup>We run the source code provided by [38] to obtain the same object candidates.

Method	CAD alignment GT	Manual GT
VDPM-16 [42]	NA	51.9
Xiang et al. [28]	64.4	64.3
Random CAD [42]	NA	61.8
GT CAD [42]	NA	67.3
DISCO	<b>71.2</b>	<b>67.6</b>

Table 3: Object segmentation accuracies (%) of different methods on PASCAL3D+. Best results are shown in red.

Method	Sofa		Chair	
	Avg. Recall	PCK	Avg. Recall	PCK
3D-INN	<b>88.0</b>	31.0	87.8	41.4
DISCO	83.4	<b>38.5</b>	<b>89.9</b>	<b>63.9</b>

Table 4: Average recall and PCK[ $\alpha = 0.1$ ] accuracy(%) for 3D structure prediction on the sofa and chair classes in IKEA dataset.

segmentation accuracy<sup>4</sup> of its projected model mask. With a 3D skeleton of an object, we are able to create a coarse object mesh based on the geometry and compute segmentation masks by projecting coarse mesh surfaces onto 2D image based on the estimated 2D keypoint locations. Please refer to the supplementary document for more details.

Table 3 reports the object segmentation accuracies on two types of ground truths. The column “Manual GT”, is the manual pixel-level annotation provided by PASCAL VOC 2012, whereas “CAD alignment GT” uses the 2D projections of aligned CAD models as ground truth. Note that “CAD alignment GT” covers the entire object extent in the image including regions occluded by other objects. DISCO significantly outperforms the state-of-the-art method [41] by 4.6% and 6.6% using only synthetic data for training. Moreover, on “Manual GT” benchmark, we compare DISCO with “Random CAD” and “GT CAD” which stand for the projected segmentation of randomly selected and ground truth CAD models respectively, given the ground truth object pose. We find that DISCO yields even superior performance to “GT CAD”. This provides evidence that joint modeling of 3D geometry manifold and viewpoint is better than the pipeline of object retrieval plus alignment. Further, we would like to emphasize much faster inference of a single forward pass of DISCO during testing compared with other sophisticated CAD alignment approaches.

#### 4.4. IKEA Dataset

In this section, we evaluate DISCO on IKEA dataset [20] with 3D keypoint annotations provided by [39]. We train a single DISCO from scratch using 200K synthetic images of both chair and sofa instances, in order to evaluate whether DISCO is capable of learning multiple 3D object geometries simultaneously. At test time, we compare DISCO

with the state-of-the-art 3D-INN[39] on IKEA. In order to remove the error on the viewpoint estimation for 3D structure evaluation as 3D-INN does, we compute the PCA bases of both the estimated 3D keypoints and their groundtruth. Next, we align two PCA bases and rotate the predicted 3D structure back to the canonical frame of the groundtruth. Table 4 reports the PCK[ $\alpha = 0.1$ ] and average recall[39] (mean PCK over densely sampled  $\alpha$  within  $[0, 1]$ ) of 3D-INN and DISCO on both sofa and chair classes. We retrieve the PCK accuracy for 3D-INN from its publicly released results on IKEA dataset. DISCO significantly outperforms 3D-INN on PCK, which means that DISCO obtains more correct predictions than 3D-INN. This substantiates that direct exploitation of the rich visual details from images adopted by DISCO is critical to infer more accurate and fine-grained 3D structure than building 3D based on sparse 2D keypoints like in 3D-INN. However, DISCO is inferior to 3D-INN in terms of average recall on the sofa class. This indicates that the wrong predictions by DISCO deviates more from the groundtruth than 3D-INN. This is mainly because 3D predicted shapes from 3D-INN are constrained by shape bases so that the wrong estimations still maintain rough object shapes when recognition fails. We conclude that DISCO is able to learn 3D patterns of object classes besides the car category and shows potential as a general-purpose approach to jointly model 3D geometry structures of multiple objects.

#### 4.5. Qualitative Results

In Figure 4, we demonstrate example predictions from DISCO on KITTI-3D (left column) and PASCAL VOC (right column). From left to right, each row shows the original object image, the predicted 2D object skeleton as well as instance segmentation and 3D object skeleton with visibility.

We show 2D and 3D structure prediction results under no occlusion (rows 1), truncation (row 2), multi-car occlusion (row 3) and other occluders (row 4). Note that DISCO can localize 2D and 3D keypoints on real images with complex occlusion scenarios and diverse car models such as sedan, SUV and pickup. Moreover, the visibility inference by DISCO is mostly correct. These capabilities highlight the potential of DISCO as a building block towards holistic scene understanding in cluttered scenes. The last row shows two failure cases where the left car is mostly occluded by another object and the right one is severely truncated and distorted in projection. We may improve the performance of DISCO on these challenging cases by training DISCO on both synthetic data simulated with more complex occlusions [30] and real data with 2D and 3D annotations.

Finally, we qualitatively compare 3D-INN and DISCO on two examples from IKEA dataset visualized in Fig. 5. In the chair example, 3D-INN fails to delineate the inclined seat-back. For the sofa, DISCO captures the sofa armrest whereas

<sup>4</sup>The standard IoU segmentation metric on PASCAL VOC benchmark.



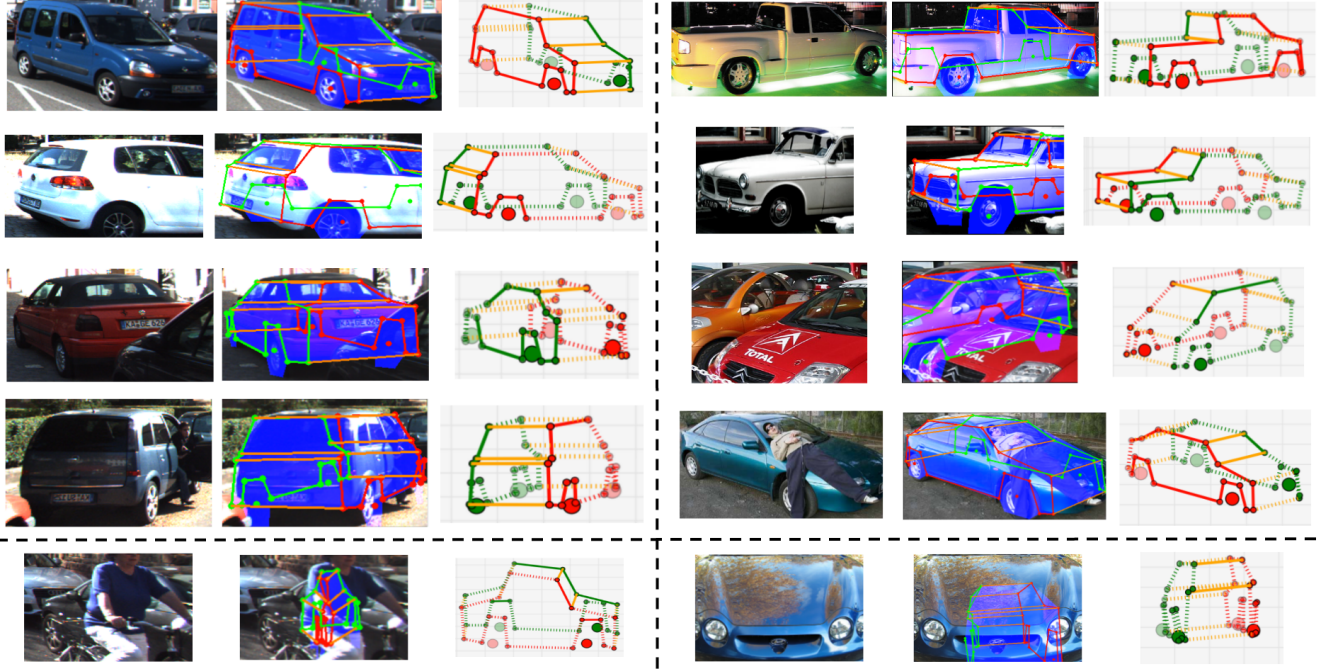


Figure 4: Visualization of 2D/3D prediction, visibility inference and instance segmentation on KITTI-3D (left column) and PASCAL VOC (right column). Last row shows failure cases. Circles and lines represent keypoints and their connections. Red and green indicate the left and right sides of a car, orange lines connect two sides. Dashed lines connect keypoints if one of them is inferred to be occluded. Light blue masks present segmentation results.

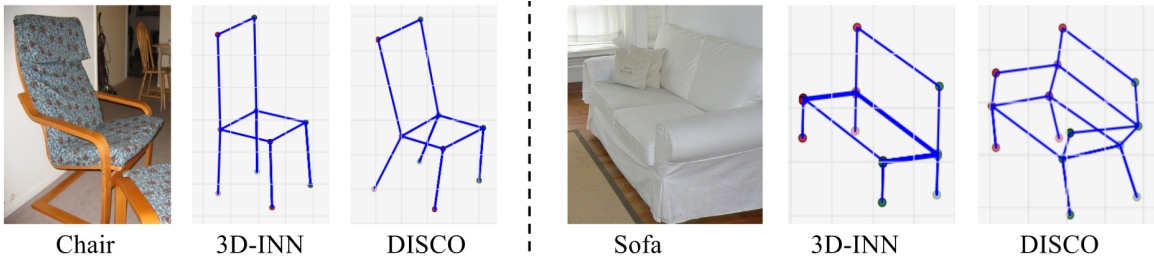


Figure 5: Qualitative comparison between 3D-INN and DISCO for 3D structure prediction on IKEA dataset.

3D-INN merges armrests to the seating area. We attribute this success to the direct mapping from image evidence to 3D structure, as opposed to lifting 2D keypoint predictions to 3D.

## 5. Conclusion

We present a framework that deeply supervises a CNN architecture to incrementally develop 2D/3D shape understanding using a series of intermediate shape concepts. A 3D CAD model rendering pipeline generates numerous synthetic training images with supervisory signals for the deep supervision. The fundamental relationship of the shape concepts to 3D reconstruction is confirmed by the fact that our network generalizes well to real images at test time, despite

synthetic data not satisfying certain aspects of photorealism. Generalization of CNNs from synthetic data to real images has been considered in recent works [6, 25, 31], but our approach goes further in providing explicit supervision signals through intermediate shape concepts. We postulate that this lends richer insights while reducing over-fitting in the learning process. Experiments demonstrate that our network outperforms current state-of-the-art methods on 2D and 3D landmark prediction on public datasets, even with occlusion and truncation. Further, we present preliminary results on jointly learning 3D geometry of multiple object classes within one single CNN. Our future work will extend this direction by learning representations for diverse object classes. The present method is unable to model highly deformable objects due to lack of CAD training data, and topologically



inconsistent object categories such as buildings, which is another focus for the future. More interestingly, our deep supervision can be potentially applied to tasks with abundant intermediate concepts such as scene physics inference.

## References

- [1] M. Aubry, D. Maturana, A. Efros, B. Russell, and J. Sivic. Seeing 3D chairs: Exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *CVPR*, 2014. [2](#)
- [2] A. Bansal, B. Russell, and A. Gupta. Marr revisited: 2D-3D Alignment via Surface Normal Prediction. In *CVPR*, 2016. [2](#)
- [3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015. [4](#), [5](#)
- [4] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction. In *ECCV*, 2016. [2](#)
- [5] A. Dosovitskiy, J. Springenberg, and T. Brox. Learning to Generate Chairs with Convolutional Neural Networks. In *CVPR*, 2015. [2](#)
- [6] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. [8](#)
- [7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012. [2](#), [5](#)
- [8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. [6](#)
- [9] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. *AISTATS*, 2010. [5](#)
- [10] S. Gupta, P. Arbeláez, R. Girshick, and J. Malik. Inferring 3d object pose in RGB-D images. *arXiv:1502.04652*, 2015. [2](#)
- [11] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *JMLR*, 2015. [4](#)
- [12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. [5](#)
- [13] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly Supervised Matching for Single-view Reconstruction. In *CVPR*, 2016. [3](#), [5](#), [6](#)
- [14] A. Kar, S. Tulsiani, J. Carreira, and J. Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. [2](#)
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. [4](#)
- [16] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015. [2](#)
- [17] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu. Deeply-Supervised Nets. *AISTATS*, 2015. [2](#), [3](#)
- [18] H.-J. Lee and Z. Chen. Determination of 3D human body postures from a single view. *CVGIP*, 1985. [2](#)
- [19] J. J. Lim, A. Khosla, and A. Torralba. FPM: Fine pose Parts-based Model with 3D CAD models. In *ECCV*, 2014. [2](#)
- [20] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing IKEA Objects: Fine Pose Estimation. In *ICCV*, 2013. [2](#), [5](#), [7](#)
- [21] J. L. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In *NIPS*, 2014. [6](#)
- [22] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publishers, Norwell, MA, USA, 1985. [1](#)
- [23] D. Marr. *Vision*. Henry Holt and Co., Inc., 1982. [1](#)
- [24] F. Massa, B. Russell, and M. Aubry. Deep Exemplar 2D-3D Detection by Adapting from Real to Rendered Views. In *CVPR*, 2015. [2](#)
- [25] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow and Scene Flow Estimation. In *CVPR*, 2016. [8](#)
- [26] R. Mohan and R. Nevatia. Using perceptual organization to extract 3D structures. *PAMI*, 1989. [1](#)
- [27] P. Moreno, C. K. Williams, C. Nash, and P. Kohli. Overcoming occlusion with inverse graphics. In *ECCV*, 2016. [2](#)
- [28] R. Mottaghi, Y. Xiang, and S. Savarese. A coarse-to-fine model for 3d pose estimation and sub-category recognition. In *CVPR*, 2015. [7](#)
- [29] D. J. Rezende, S. Eslami, S. Mohamed, P. Battaglia, M. Jaderberg, and N. Heess. Unsupervised learning of 3d structure from images. In *NIPS*, 2016. [2](#)
- [30] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. [7](#)
- [31] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. [8](#)
- [32] S. Sarkar and P. Soundararajan. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata. *PAMI*, 2000. [1](#)
- [33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. [4](#)
- [34] B. J. Smith. *Perception of Organization in a Random Stimulus*. 1986. [1](#)
- [35] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with Rendered 3D model views. In *ICCV*, 2015. [2](#), [4](#)
- [36] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3D Models from Single Images with a Convolutional Network. In *ECCV*, 2016. [2](#)
- [37] L. Torresani, A. Hertzmann, and C. Bregler. Learning non-rigid 3d shape from 2d motion. In *Advances in Neural Information Processing Systems*, page None, 2003. [1](#)
- [38] S. Tulsiani and J. Malik. Viewpoints and Keypoints. In *CVPR*, 2015. [2](#), [6](#)
- [39] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single Image 3D Interpreter Network. In *ECCV*, 2016. [2](#), [5](#), [7](#)
- [40] T. Wu, B. Li, and S.-C. Zhu. Learning And-Or Model to Represent Context and Occlusion for Car Detection and Viewpoint Estimation. *PAMI*, 2016. [2](#)

- [41] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3D voxel patterns for object category recognition. In *CVPR*, 2015. 2, 7
- [42] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A Benchmark for 3D Object Detection in the Wild. In *WACV*, 2014. 2, 5, 6, 7
- [43] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 5, 6
- [44] X. Yu, F. Zhou, and M. Chandraker. Deep Deformation Network for Object Landmark Localization. *ECCV*, 2016. 3, 5, 6
- [45] T. Zhou, P. Krähenbühl, M. Aubry, Q. Huang, and A. A. Efros. Learning Dense Correspondence via 3D-guided Cycle Consistency. In *CVPR*, 2016. 2, 3
- [46] M. Z. Zia, U. Klank, and M. Beetz. Acquisition of a Dense 3D Model Database for Robotic Vision. In *ICAR*, 2009. 2
- [47] M. Z. Zia, M. Stark, and K. Schindler. Towards Scene Understanding with Detailed 3D Object Representations. *IJCV*, 2015. 2, 5, 6